

一种嵌入项目疲劳和多样偏好的聚合推荐算法 *

阙正昊, 邓明通, 刘学军, 李 斌

(南京工业大学 计算机科学与技术学院, 南京 211816)

摘 要: 为了解决推荐列表偏向于热门项目, 多样性差的问题, 提出了 ARIFDP 算法 (aggregation recommendation algorithm for embedding item fatigue and diversity preference)。首先通过对用户历史反馈数据分析用户的多样性偏好, 得出用户的多样倾向度, 进而构造了与评价次数负相关的项目疲劳函数, 最终将矩阵分解与项目疲劳函数相聚合, 并加入多样倾向度调节项目疲劳函数所占权重, 增加了冷门项目被推荐的概率。实验结果表明, ARIFDP 算法能在保证准确率的前提下有效提高推荐结果的多样性。

关键词: 主题模型; 矩阵分解; 多样倾向度; 项目疲劳函数; 推荐多样性

中图分类号: TP301.6 doi: 10.3969/j.issn.1001-3695.2018.04.0279

Aggregation recommendation algorithm for embedding item fatigue and diversity preference

Que Zhenghao, Deng Mingtong, Liu Xuejun, Li Bin

(College of Computer Science & Technology, Nanjing Tech University, Nanjing 211816, China)

Abstract: In order to solve the problem that the recommendation list is biased towards popular projects and with poor diversity, this paper proposes the ARIFDP (An Aggregation Recommendation Algorithm for Embedding Item Fatigue and Diversity Preference) algorithm. First, the user's diversity preferences are analyzed by using the historical feedback data of the user to derive the user. The degree of diversification; and then constructing a project fatigue function negatively related to the number of evaluations; eventually the matrix decomposition and the project fatigue function are aggregated, and the inclusion of various propensity degrees adjusts the weight of the fatigue function of the project, and increases the probability of the proposed project. Experimental results show that the ARIFDP algorithm can effectively improve the diversity of recommendation results on the premise of ensuring accuracy.

Key words: topic model; matrix factorization; diversity tendency; item fatigue function; recommendation diversity

0 引言

近年来, 推荐系统已经成为电子商务网站或资料库网站中帮助其推广业务以及帮助用户寻找项目的必备工具。作为一个有效的解决信息过载问题的工具, 推荐系统能够从巨大的项目池中给用户推荐符合其偏好的项目集合。

目前已有的推荐系统多注重准确性, 如何提高多样性越来越成为推荐任务中的关键问题^[1]。例如, 当用户在电影资料库网站 IMDB(<http://www.imdb.com/title/tt3612032/>) 中搜索电影《致命弯道》后, 所有关于《致命弯道》系列的电影都会被推荐, 如图 1 所示。从准确性的角度, 推荐列表是令人满意的, 因为目标用户喜欢《致命弯道》。然而, 由于推荐结果缺乏多样性, 不利于拓宽用户的视野, 从而让用户感到厌烦。事实上,

拓宽用户的视野已经成为推荐系统的重要特性之一。^{错误!未找到引用源。}

。一个能够开拓用户视野的系统可以获得一个双赢的结果: 用户可以找到更多的有趣的项目, 网站经营者可以增加他们的销售额, 提高用户满意度。本文通过对用户的历史反馈信息分析得出用户的多样倾向度, 将用户的历史反馈信息和项目的主题分布使用矩阵分解得到用户对项目的初始偏好得分, 进而通过项目疲劳函数调整偏好得分, 并加入多样倾向度调节项目疲劳函数所占比重, 从而向用户推荐既具有多样性, 同时也不失准确性的一组推荐列表。

1 相关工作

推荐系统的主要目标是向用户推荐一组满足其要求的个性化列表, 根据用户的历史行为数据做出推荐, 文献^{错误!未找到}

收稿日期: 2018-04-09; 修回日期: 2018-06-11 基金项目: 江苏省重点研发计划 (社会发展) 资助项目 (BE2015697); 国家自然科学基金资助项目 (61203072)

作者简介: 阙正昊 (1991-), 男, 江苏淮安人, 硕士研究生, 主要研究方向为数据挖掘、推荐系统等 (450017507@qq.com); 邓明通 (1991-), 男, 硕士研究生, 主要研究方向为数据挖掘、推荐系统等; 刘学军 (1970-), 男, 教授, 博士, 主要研究方向为数据库、数据挖掘、传感器网络等; 李斌 (1979-), 男, 硕士, 讲师, 主要研究方向为传感器网络、智能信息处理等。

引用源。针对用户的购买或点击行为, 提出了三种折扣方法: 面向用户的折扣, 面向项目的折扣和面向时间的折扣, 该方法能够很好的与现有的矩阵分解模型相结合, 提高了推荐的准确性。文献**错误!未找到引用源。**认为需要同时收集其正反馈和负反馈数据, 提出了将用户反馈作为分类变量, 并且将它和用户, 项目以三元组的方式建模, 采用了三阶张量分解技术和更高阶的折叠式方法产生推荐。文献**错误!未找到引用源。**提出了一种基于一组不同属性(包括认知努力, 用户模型, 测量尺度和域相关性), 在推荐系统中使用显式和隐式用户反馈的分类框架, 提高了推荐系统的性能。上述的推荐方法虽然一定程度上提高了推荐准确性, 但是都存在一定的不足之处, 即忽略了推荐的多样性。

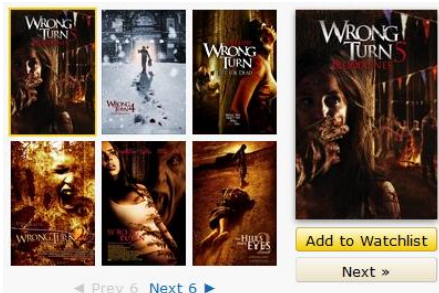


图1 IMDB 网站中关于《致命弯道》的推荐列表

目前对于多样性的研究主要分为两个方面: 总体多样性 (aggregate diversity) 和个体多样性 (individual diversity)。文献**错误!未找到引用源。**利用异构信息网络中的元路径相似性度量计算得到项目间的相关性, 从而关联流行项目和利基项目, 提高了推荐的总体多样性。文献**错误!未找到引用源。**为了提高推荐的个体多样性, 提出了使用张量分解揭示包括社区、用户和社会标签的多模式数据中潜在主题的框架。文献**错误!未找到引用源。**提出了一种面向多样性的热传导算法, 与面向准确度的能量扩散方法结合, 调节精确度与多样性的平衡。文献**错误!未找到引用源。**结合了传统协同过滤和概率主题建模的优点, 为用户和项目提供了可解释的潜在结构, 增加了推荐的多样性。文献**错误!未找到引用源。**为了发现新颖项目, 提出了通过项目的共现性将用户和项目之间的关系转换为项目和项目之间的关系 UC-BCF 模型。上述学者对于多样性的研究都集中在项目集合上, 而没有考虑到用户自身对于多样性的偏好程度。

本文提出了嵌入项目疲劳和多样偏好的聚合推荐算法, 主要致力于向目标用户推荐符合其多样性偏好程度的个性化列表。本文的主要工作如下:

- 通过对用户历史反馈信息分析用户对于多样性的偏好程度, 提出了多样倾向度的概念, 并通过聚类分析计算出用户的多样倾向度。
- 为了增加冷门项目被推荐的权重, 通过分析项目的被浏览信息, 构造了受长尾分布约束的项目疲劳函数。
- 提出了对用户历史反馈信息和项目主题信息使用矩阵分解与项目疲劳函数相聚合的 ARIFDP 算法, 并加入多样倾向度调节项目疲劳函数所占权重。

d) 在真实数据集上对本文提出的方法进行了实验, 结果证实 ARIFDP 能够获得较好的结果。

2 问题描述和 LDA 主题模型

2.1 问题描述

本文按如下方式考虑推荐问题: 给定项目的描述信息以及目标用户的历史反馈信息, 向目标用户推荐符合其多样性偏好程度的项目集合。

2.2 LDA 主题模型

LDA 模型**错误!未找到引用源。**由 D.M.Blei 在 2003 年提出, 是一种非监督机器学习技术, 可以用来识别大规模文档集(document collection)或语料库(corpus)中潜藏的主题信息。其模型图如图 2 所示。

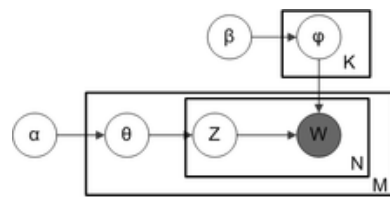


图2 LDA 模型图

当给定一个有 M 篇文档的文档集合 D , 共包含 K 个主题 z , N 个单词 w 。其中 α 与 β 是服从狄利克雷分布的语料级别的参数, α 是 $p(\theta)$ 的向量参数, 用于生成一个主题 θ 向量, β 是各个主题对应的单词概率分布矩阵 $p(w|z)$ 。则文本的生成过程可描述如下:

- 对每个文档 $d \in D$, 从狄利克雷分布 $Dir(\alpha)$ 中取样生成文档 d 的主题分布 θ ;
- 从主题的多项式分布 θ 中取样生成文档 d 第 n 个单词的主题 z_n ;
- 从狄利克雷分布 $Dir(\beta)$ 中取样生成主题 z_n 的词分布 φ_{z_n} ;
- 从词的多项式分布 φ_{z_n} 中采样最终生成词 w_n 。

M 个文档集合 D 用 LDA 生成的概率为

$$\prod_{m=1}^M \int p(\theta_m | \alpha) \left(\prod_{n=1}^{N_m} p(z_n | \theta_m) p(\varphi_{z_n} | \beta) p(w_n | \varphi_{z_n}) \right) d\theta_m \quad (1)$$

上述表达式中含有两个隐含变量, 即文档-主题分布 θ 和主题-词分布, 它们均可由 Gibbs sampling 方法获得。

3 嵌入项目疲劳和多样偏好的聚合推荐算法

3.1 用户多样性偏好

定义 1 多样倾向度。设项目集合中所有项目经过聚类后的类簇数为 K , 用户反馈信息中项目的子类数目为 L , 则用户的多样倾向度为 $\omega = L / K$ 。

定义 2 项目主题分布**错误!未找到引用源。**对于每个项目 v_j , 用 LDA 主题模型获取项目描述信息的主題分布, 亦即项目主题分布, 记为 θ_v 。

通过对用户历史反馈信息研究发现, 大部分用户的历史反

馈信息中都存在很多不同种类的项目, 项目的种类数越多则说明用户越偏好于多样性。本文通过 K 均值聚类算法对项目聚类, K 均值聚类算法是一种基于形心的技术, 需要计算不同类型的项目相异度。对于项目特征, 可以通过项目的描述信息或项目的属性信息获取, 本文通过 LDA 主题模型提取每个项目描述信息的主题分布向量, 通过欧几里得距离计算每两个向量 a 和向量 b 之间的相异度 $dist(a, b)$ 为

$$dist(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2} \quad (2)$$

其中: n 为主题向量的维数。在设定类簇个数 K 并完成聚类后, 本文通过平均轮廓系数^{错误!未找到引用源。}进行有效性分析, 对象 o 的轮廓系数 $s(o)$ 为

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}} \quad (3)$$

在 K -均值聚类中, $s(o)$ 的值在 0 和 1 之间, $a(o)$ 表示 o 与 o 所属的簇的其他对象之间的平均距离, 其值反映 o 所属的簇的紧凑性, 该值越小, 簇越紧凑; $b(o)$ 表示 o 与不属于 o 的所有簇的最小平均距离, 其值捕获 o 与其他簇的分离程度, 该值越大, o 与其他簇越分离。为了度量聚类的质量, 求所有项目的轮廓系数的平均值:

$$\overline{s(o)} = \frac{1}{n} \sum_{i=1}^n s(o) \quad (4)$$

其中: n 为项目总数。平均轮廓系数的值越大, 聚类质量越好。对于所有可能的类簇个数 K , 求取平均轮廓系数的最大值, 此时的 K 即为最佳的聚类簇数。

使用 K -均值聚类算法得到所有项目的 K 个类簇后(离群点除外), 将用户的历史反馈记录中的项目逐一与聚类完成后的 K 个簇进行比对后可以得到 L 个子类 ($L \leq K$), L 即为用户历史反馈记录中项目的子类数量。则用户的多样倾向度通过定义 1 即可求得。

算法 1 多样倾向度取值算法

输入: 项目集合中所有项目描述信息的数据集 S_1 , 用户历史反馈数据中项目 ID 的数据集 S_2 。

输出: 多样倾向度 ω 。

begin

a) for each $i \in S_1$

b) 采用 LDA 计算项目描述信息的主题分布向量 θ

c) end for

d) 分别计算每两个向量之间的相异度 $dist(a, b)$

e) 从 S_1 中任意选择 K 个对象作为初始簇中心

f) repeat

g) 根据簇中对象的均值, 将每个对象分配到最相似的簇

h) 更新簇均值, until 各个簇均值不再发生变化

i) 变换 K 的数值, 重复步骤 5~8, 分别求其平均轮廓系数, 取最大平均轮廓系数对应的 K 值作为 S_1 聚类后项目的类簇数 K

j) for each $j \in S_2$

k) 将 j 与 K 个簇进行比对

l) end for

m) 计算得到 S_2 中共有 L 个子类

n) $\omega \leftarrow L / K$

o) return ω

end

3.2 多样性推荐方法

传统的协同过滤方法能够解决一定的稀疏性问题, 但是当有新项目加入到系统中时, 没有足够的用户对其产生过反馈信息, 此时如果利用传统的协同过滤方法就很难得到理想的推荐效果。受文献^{错误!未找到引用源。}的启发, 将项目的主题模型融入矩阵分解中计算出用户对项目的初始偏好得分, 同时聚合项目疲劳函数, 调整偏好得分, 一方面可以缓解冷启动问题, 另一方面也可以增加推荐结果的多样性。

为了改善矩阵分解的推荐效果, 增加冷门项目的信息, 本文用 LDA 主题模型学习项目描述信息的主题分布 θ_v , 不采用随机的方式初始化项目的隐含特征矩阵, 而是用 θ_v 初始化项目的隐含特征矩阵 V , 通过模型学习得到的项目的隐含特征 V 如下:

$$v_j = \varepsilon_{vj} + \theta_{vj} \quad (5)$$

其中: ε 表示项目的特征分布与 LDA 学习得出的主题分布的偏差值, 项目的特征分布是接近于 LDA 学习出的主题分布, 但也存在偏差。LDA 使用的是内容信息获取项目的主题分布, 在 LDA 与矩阵分解的融合中加入用户的评分信息, 调整用户和项目的特征分布, 则目标函数如下:

$$f(U, V) = \sum_{(i, j) \in I} (r_{ij} - u_i^T v_j)^2 + \lambda_u \sum_i \|u_i\|^2 + \lambda_v \sum_j \|v_j - \theta_{vj}\|^2 \quad (6)$$

对于每一个用户 i 和项目 j 的对 (i, j) , r_{ij} 表示用户 i 对项目 j 的评分。 $r_{ij} \in R$, R 为评分矩阵。 λ_u 和 λ_v 是正则项系数, 通过使用交替最小二乘法求解可以得到 u_i 和 v_j 相应的更新公式:

$$u_i = (VV^T + \lambda_u E)^{-1} V R_i \quad (7)$$

$$v_j = (UU^T + \lambda_v E)^{-1} (U R_j + \lambda_v \theta_{vj}) \quad (8)$$

其中: E 是单位矩阵; R_i 是 R 矩阵中第 i 行向量的转置; R_j 是 R 矩阵中第 j 行向量的转置。通过迭代更新 u_i 和 v_j 可以得到用户隐含特征 U^* 和项目隐含特征 V^* , 则用户 i 对项目 j 的偏好得分 r_{ij}^* 计算公式如下:

$$r_{ij}^* = (u_i^*)^T v_j^* \quad (9)$$

对电子商务网站的销售记录研究发现, 项目的销量呈长尾分布, 项目的被评价次数也同样呈长尾分布^{错误!未找到引用源。}, 当用

户被多次推荐相同的项目会产生项目疲劳^[14]。为了进一步增加冷门项目被推荐的概率, 本文将矩阵分解计算得出的初始偏好得分与项目疲劳函数相聚合, 提高冷门项目的推荐权重。因此将式(9)改写为

$$r_{ij}^* = (u_i^*)^T v_j^* + \omega \cdot \mu \eta(x) \quad (10)$$

其中: $\eta(x)$ 为项目疲劳函数, 其具体构造方法将在下一节详细介绍; ω 为多样倾向度, 用来调节项目疲劳函数所占比重; 系数 μ 用来调节 $\eta(x)$ 的值域范围, 其值为评分范围的最大值。

将对应的偏好得分降序排序, 选取 top- N 作为最终推荐列表。

3.3 受长尾分布约束的项目疲劳函数

定义 3 长尾分布^{错误!未找到引用源。}。令 $S(x)$ 为任意一个分布的累积分布函数, 互补函数为 $S^c(x) = 1 - S(x)$ 。如果满足对任意 $\gamma > 0$, 当 $t \rightarrow \infty$ 时, 有 $e^{\gamma t} S^c(x) \rightarrow \infty$, 则称 $S(x)$ 对应的分布为长尾分布。

为了满足定义 3 中描述的长尾分布的条件, 文献^{错误!未找到引用源。}提出了由 n 个底为 e 的指数函数线性组合描述长尾分布函数:

$$P(x) = \sum_{j=1}^n p_j e^{-\lambda_j x} + C \quad (11)$$

数据集中项目的被评价次数符合长尾分布, 将数据点记为 (x_i, y_i) ($i=1, \dots, h$), 其中, x_i 为项目被评分次数的排名, y_i 为项目被评分的次数, h 为项目总数, $F(x)$ 表示 $P(x)$ 与数据 (x_i, y_i) 的平方差:

$$F(x) = \sum_{i=1}^h \omega_i \left(\sum_{j=1}^n p_j e^{-\lambda_j x_i} + C - y_i \right)^2 \quad (12)$$

其中: $\omega_i > 0$ 为点的权系数, 可通过求 n 阶非线性方程组得到使 $F(x)$ 最小的参数 $p_j, \lambda_j > 0$ ($j=1, \dots, n$), 进而求得满足长尾分布描述的 $P(x)$ 。文献^{错误!未找到引用源。}进一步通过对韦伯分布的拟合效果分析得出当 $n=2$ 时, 对长尾分布的描述效果最优, 也即 $P(x) = p_1 e^{-\lambda_1 x} + p_2 e^{-\lambda_2 x} + C$ 。

由于项目被评价的次数呈长尾分布, 为了增加对冷门项目的推荐, 提高推荐的多样性, 通过构造与评价次数负相关的项目疲劳函数来提高冷门项目的推荐权重, 令函数 $g(x) = 1/(1+e^{-x})$ 将长尾分布的函数值 $P(x)$ 映射到 $[0,1]$, 则项目疲劳函数 $\eta(x)$ 可表示如下:

$$\eta(x) = 1 - \frac{1}{1 + e^{-P(x)}} \quad (13)$$

3.4 ARIFDP 算法总体描述

推荐系统要以“用户的长期维系”为目标, 如果只以当前大多数基于准确率为衡量标准的算法进行推荐, 将严重影响推荐系统的长期性能, 从而导致长尾项目无法得到推荐。本文 ARIFDP 推荐算法的主要步骤为: 通过对用户历史反馈信息分

析用户的多样倾向度; 将主题模型融入矩阵分解计算出用户对项目的初始偏好得分; 为了提高推荐多样性, 将初始偏好得分与项目疲劳函数相聚合, 并通过多样倾向度调节项目疲劳函数所占权重。ARIFDP 算法的主要步骤表示如下:

算法 2 多样性推荐算法 ARIFDP

输入: 用户评分数据集 D , 待推荐项目集合 item。

输出: 每个用户的 top- N 推荐列表。

begin

1) for each $i \in Item$

2) 通过项目的描述信息计算项目主题分布 θ_v

3) end for

4) 用 θ_v 初始化项目隐含特征矩阵 V , 并加入用户评分数据集 D 调整用户和项目的特征分布, 求解 U^* 和 V^*

5) 计算用户对于项目的初始偏好得分 $r_{ij}^* \leftarrow (u_i^*)^T v_j^*$

6) 通过数据集 D 构造项目疲劳函数 $\eta(x) \leftarrow 1 - \frac{1}{1 + e^{-W(x)}}$

7) $r_{ij}^* \leftarrow (u_i^*)^T v_j^* + \omega \cdot \mu \eta(x)$

8) 根据 r_{ij}^* 降序排序, 选取 Top- n 作为最终的推荐列表

end

4 实验及结果分析

4.1 实验数据集

本文实验数据采用了 MovieLens 数据集, 该数据集由美国明尼苏达州立大学 GroupLens 研究小组提供, 包括 6 040 个用户对 3 952 部电影的约 100 万条电影评分信息, 每个用户至少评价了 20 部电影, 且评分范围为 1~5, “1”表示“poor”(不喜欢), “5”表示“perfect”(非常喜欢)。本文从中选取了 943 个用户对 1682 部电影的大约 100 000 次评分数据作为实验数据集, 稀疏度为 93.7%。由于数据集中不包含电影的描述信息, 但是此部分信息是本文中算法的重要组成部分, 而 MovieLens 数据集中的 u.item 文件中包含了每部电影链接 IMDb URL, 因此通过 Python 爬虫程序获取每部电影的 Storyline(描述信息)。

在实验中, 从数据集中随机抽取 80% 作为训练集, 其余的 20% 作为测试集。

4.2 评价标准

4.2.1 推荐准确性评测指标

本实验采用平均绝对偏差 MAE(mean absolute error)作为度量准确性的标准, 它是一种统计精度度量方法, 同时也是最常用的一种推荐质量度量方法, 其通过计算预测评分与实际评分之间的偏差来衡量预测的准确性, MAE 的值越小, 表明推荐质量越高。

假设预测的用户评分集合为 $\{p_1, p_2, \dots, p_N\}$, 对应的实际用户评分集合为 $\{q_1, q_2, \dots, q_N\}$, N 表示预测的次数, 则 MAE 的计

算公式如下:

$$MAE = \frac{\sum_{i=1}^N |p_i - q_i|}{N} \quad (14)$$

4.2.2 推荐多样性评测指标

高准确性一直是目前大多数推荐系统的衡量标准,但是系统要以“用户的长期维系”为目标,因此提高推荐结果的多样性非常重要。对推荐结果的多样性评测主要包括两个方面:总体多样性和个体多样性。

对于总体多样性,本实验采用信息熵^{错误:未找到引用源。}进行评测。信息熵是信息理论中用于度量信息量的一个概念,如果系统越有序,那么信息熵越低,反之越高。如果推荐结果中所有的项目都能够出现,并且出现的次数相差不大,那么系统具有很好的挖掘长尾项目的能力,我们将不同推荐列表中项目的出现次数称为该项目在推荐结果中的流行度,信息熵的计算公式如下:

$$H = -\sum_{i=1}^n p(i) \log p(i) \quad (15)$$

其中: $p(i)$ 等于项目 i 的流行度除以所有项目流行度之和。信息熵越大,表明算法挖掘长尾项目的能力越好。

对于个体多样性,通过计算每个用户 u 的推荐结果 $R(u)$ 中两两项目的不相似程度 $D(R(u))$,进而求得所有用户的推荐列表不相似程度的均值为

$$D(R(u)) = 1 - \frac{\sum_{i,j \in R(u), i \neq j} \text{sim}(i,j)}{\frac{1}{2}|R(u)|(|R(u)|-1)} \quad (16)$$

$$D = \frac{1}{|U|} \sum_{u \in U} D(R(u)) \quad (17)$$

4.3 实验过程

4.3.1 类簇个数的取值

K-均值聚类算法中,不同的 K 值对聚类效果有着重要的影响,本实验采用平均轮廓系数来确定类簇的个数,数据集中共有 1682 部电影,通过分析数据集中电影的数据规模、类别特征以及已有的 K 均值算法对于 MovieLens 数据集的聚类效果,将 K 值范围设定在[2,50]之间。

由于 K-均值聚类具有一定的随机性,并不是每次都收敛到全局最小,为了避免得到局部最优解,针对区间内每一个 K 值,重复执行 30 次,计算每一次的 $\overline{s(o)}$,并取其平均值 $\text{AVG}(\overline{s(o)})$ 度量 K 值的聚类效果,由于篇幅有限,表 1 列出了部分 $\text{AVG}(\overline{s(o)})$ 较大时聚类质量情况。

表 1 不同的类簇个数对聚类效果的影响

簇数	AVG(\overline{s})	簇数	AVG(\overline{s})
8	0.5714	13	0.6912
9	0.5893	14	0.6895

10	0.5167	15	0.6713
11	0.6058	16	0.6574
12	0.6719	17	0.6051

从表 1 中可以看出,类簇数为 13 时的平均轮廓系数最大,其值为 0.6912,所以类簇数为 13 时聚类效果最好,在接下来的实验中,将类簇数设置为 13 进行实验。

4.3.2 多样倾向度 ω 的分布

在 ARIFDP 算法中,多样倾向度 ω 起到了很重要的作用,用来平衡准确性和多样性的比重, ω 越大表明算法受项目疲劳函数的影响力越大。由 4.3.1 节可知, K=13 时聚类效果最好,本实验分别计算了数据集在 K=13 时 30 次聚类完成后各用户评分过的电影类别数,并计算了用户 30 次多样倾向度 ω 的平均值,得到用户的多样倾向度 ω 的分布情况,结果如表 2 所示。

表 2 不同的 ω 值所占的百分比

ω 值的范围	所占百分比
[0,0.1)	3.5%
[0.1,0.2)	9%
[0.2,0.3)	57%
[0.3,0.4)	15%
[0.4,0.5)	6.5%
[0.5,0.6)	4.5%
[0.6,0.7)	2%
[0.7,0.8)	1.5%
[0.8,0.9)	0.5%
[0.9,1]	0.5%

从表 2 中可以看出,大部分的用户的多样倾向度 ω 的值集中在[0.2,0.3)内,这说明了大部分用户对于推荐系统的多样性是有一定的要求,如果只以传统的基于准确性为标准的算法进行推荐是不能够满足要求的,这也说明了本文算法的有效性。

4.3.3 多样性和准确性的对比实验

为了展示本文提出的 ARIFDP 算法的有效性和高效性,本文同以下两种方法进行了比较:

a) 传统 MF,该方法通过对用户的浏览行为进行矩阵分解来计算用户对于项目的偏好。

b) CTR^{错误:未找到引用源。},该方法结合了传统协同过滤和概率主题建模的优点,为用户和项目提供了可解释的潜在结构,增加了推荐的多样性。

由于 MovieLens 数据集中用户对电影的评分范围为 1~5,因此设置式 (10) 中的参数 $\mu=5$ 。实验中分别比较了本文提出的 ARIFDP 算法同其他两种算法在准确性,个体多样性以及总体多样性指标上的不同表现。以选择的推荐对象个数为横坐标,各评测指标为纵坐标,推荐对象个数从 10 开始,逐次增加,直至 100,实验结果如图 3~5 所示。

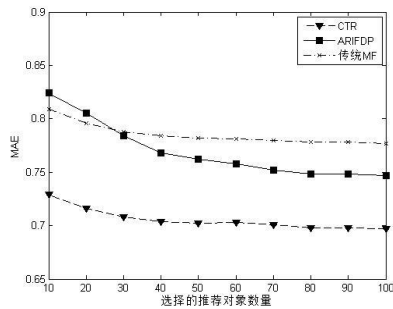


图3 不同算法的准确性对比

从图3中可以看出, 由于 ARIFDP 算法加入了项目疲劳函数, 增加了推荐的多样性, 其准确性与 CTR 模型相比略显不足, 但是由于 ARIFDP 算法不仅利用了用户的评分信息, 同时还加入了项目主题模型, 而传统的 MF 模型仅利用了用户的评分信息, 所以当推荐的项目逐渐增多的时候, 准确率较传统的 MF 模型稍高。因此, ARIFDP 算法的准确性还是在可以接受的范围之内。

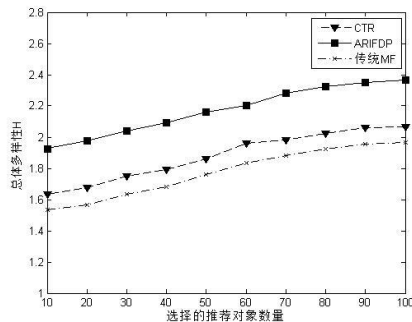


图4 不同算法的总体多样性对比

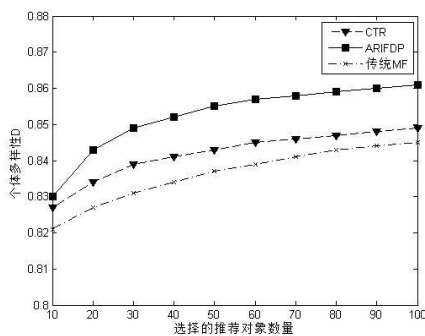


图5 不同算法的个体多样性对比

从图4和5中可以看出, 由于本文提出的 ARIFDP 算法加入了项目疲劳函数, 增加了冷门项目的推荐权重, 综合三种不同的推荐算法, 包括总体多样性和个体多样性两个方面, ARIFDP 算法均能够取得较好的效果。在总体多样性方面, ARIFDP 能够在全局上推荐更多不同种类的项目, 有利于保证推荐系统的长期性能; 在个体多样性方面, ARIFDP 能够为单个用户提供尽可能丰富的推荐列表, 有利于拓宽用户的视野, 提高用户的满意度。因此, 本文提出的 ARIFDP 算法能够有效向目标用户推荐符合其多样性偏好程度的项目集合。

5 结束语

针对如何向目标用户推荐符合其多样性偏好程度的项目集合, 本文提出了嵌入项目疲劳和多样偏好的聚合推荐算法 ARIFDP, 通过 K-均值聚类算法对项目进行聚类, 分析用户的多样性偏好程度, 得出用户的多样倾向度; 将矩阵分解与项目疲劳函数相聚合, 同时加入多样倾向度调节项目疲劳函数所占权重, 增加了冷门项目被推荐的概率。在真实数据集上的实验表明, ARIFDP 能够得到准确率较好且多样性丰富的推荐列表。

用户的多样倾向度还和用户的属性信息相关联, 比如性别、年龄、职业、受教育程度等。此外, 用户的多样倾向度是动态变化的。如何设计一个能够有效结合用户行为信息和用户属性信息的模型来度量用户的多样倾向度将是我们下一步的研究方向。

参考文献:

- [1] Cheng Peizhe, Wang Shuaiqiang, Ma Jun, *et al.* Learning to recommend accurate and diverse items [C]// Proc of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017: 183-192.
- [2] Kawai K, Kitagawa H. Collaborative filtering with implicit feedbacks by discounting positive feedbacks [C]// Proc of the 2nd IEEE International Conference on Multimedia Big Data. 2016: 41-48.
- [3] Frolov E, Oseledets I. Fifty shades of ratings: how to benefit from a negative feedback in top-N recommendations tasks [C]// Proc of the 10th ACM Conference on Recommender Systems. New York: ACM Press, 2016: 91-98.
- [4] Jawaheer G, Weller P, Kostkova P. Modeling user preferences in recommender systems: a classification framework for explicit and implicit user feedback [J]. ACM Trans on Interactive Intelligent Systems, 2014, 4 (2): 8.
- [5] Liu Yezheng, Wang Jinkun, Jiang Yuanchun, *et al.* Utilize item correlation to improve aggregate diversity for recommender systems [C]// Proc of IEEE International Conference on Data Science in Cyberspace. 2016: 412-417.
- [6] Koochi M R, Hussin A R C, Dahlan H M. Improving recommendation diversity using tensor decomposition and clustering approaches [C]// Proc of the 4th World Congress on Information and Communication Technologies. 2014: 240-245.
- [7] Zhou Tao, Kuscsik Z, Liu JianGuo, *et al.* Solving the apparent diversity-accuracy dilemma of recommender systems [J]. Proceedings of the National Academy of Sciences, 2010, 107 (10): 4511-4515.
- [8] Wang Chong, Blei D M. Collaborative topic modeling for recommending scientific articles [C]// Proc of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM Press, 2011: 448-456.

- [9] Niemann K, Wolpers M. A new collaborative filtering approach for increasing the aggregate diversity of recommender systems [C]// Proc of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2013: 955-963.
- [10] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3 (1): 993-1022.
- [11] Steinberger J. Update summarization based on novel topic distribution [C]// Proc of ACM Symposium on Document Engineering. New York: ACM Press, 2009: 205-213.
- [12] Jain R, Koronios A. Innovation in the cluster validating techniques [J]. Fuzzy Optimization & Decision Making, 2008, 7 (3): 233-241.
- [13] Park Y J, Tuzhilin A. The long tail of recommender systems and how to leverage it [C]// Proc of ACM Conference on Recommender Systems. New York: ACM Press, 2008: 11-18.
- [14] 印桂生, 张亚楠, 董红斌, 等. 一种由长尾分布约束的推荐方法 [J]. 计算机研究与发展, 2013, 50 (9): 1814-1824. (Yin Guisheng, Zhang Yanan, Dong Hongbin, *et al.* A long tail distribution constrained recommendation method [J]. Journal of Computer Research and Development, 2013, 50 (9): 1814-1824.)
- [15] Sar Shalom O, Koenigstein N, Paquet U, *et al.* Beyond collaborative filtering: the list recommendation problem [C]// Proc of International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2016: 63-72.
- [16] Zhang QianSheng, Jiang ShengYi. A note on information entropy measures for vague sets and its applications [J]. Information Sciences, 2008, 178 (21): 4184-4191.